A Novel Approach for Ontology-based Dimensionality Reduction for Web Text Document Classification

Mohamed K. Elhadad, *Khaled M. Badran,* and Gouda I. Salama
Computer department
Military Technical College
Cairo, Egypt
Moh.elhadad@mtc.edu.eg, khaledbadran@mtc.edu.eg, gisalama@mtc.edu.eg.

***Abstract*—Dimensionality reduction of feature vector size plays a vital role in enhancing the text processing capabilities; it aims in reducing the size of the feature vector used in the mining tasks (classification, clustering... etc.). This paper proposes an efficient approach to be used in reducing the size of the feature vector for web text document classification process. This approach is based on using WordNet ontology, utilizing the benefit of its hierarchal structure, to eliminate words from the generated feature vector that has no relation with any of WordNet lexical categories; this leads to the reduction of the feature vector size without losing information on the text. For mining tasks, the Vector Space Model (VSM) is used to represent text documents and the Term Frequency Inverse Document Frequency (TFIDF) is used as a term weighting method. The proposed ontology based approach was evaluated against the Principal component analysis (PCA) approach using several experiments. The experimental results reveal the effectiveness of our proposed approach against other traditional approaches to achieve a better classification accuracy, F-measure, precision, and recall**

Keywords— Dimensionality reduction; Principal component analysis; Feature Extraction; Feature Selection; WordNet; Ontology; web text documents classification; Semantic similarity; Vector Space Model; Term Frequency Inverse Document Frequency.

I. Introduction

A document is represented by a set of words that expresses its global meaning. In traditional approaches, a document is represented by a group of words describing its contents. Semantic approaches aim to give meaning to the terms of the document to address the shortcomings of traditional indexing based on single words [1].

With the increasing availability of web text documents, approximately 80% of the information of is organized and stored in unstructured textual format, and the rapid growth of the World Wide Web makes the task of automatic classification of text documents to become an interesting area for research. It is considered to be the key method for organizing the information, knowledge and trend detection [2].For Text documents to be classified, they are being processed and transformed from the full text version to a document vector, which makes the handling of them much easier and reducing their complexity.

This transformation is an important aspect in documents classification as it denotes the mapping of a document into a compact form of its content. The main problem with text documents classification is not only the extremely high dimensionality of text data, so the number of potential features often exceeds the number of training documents, but also the ignorance of the semantic information in them [3].

So, Dimensionality reduction of feature vector size is mandatory step for enhancing the classification process. Generally Dimensionality reduction algorithms are classified into feature extraction and feature selection algorithms [4]. Feature extraction algorithms aims in reducing the high dimensionality of the feature vector into a lower dimensional space through algebraic transformations by creating new features based on the original feature set. Classical feature extraction algorithms can be categorized into linear and nonlinear algorithms [5]. These algorithms project the data by transformations according to some optimization criterion. The most important technique is the principal component analysis (PCA) [6], which produces new attributes as linear combinations of the original variables.

On the other hand, Feature selection algorithms aims in reducing the high dimensionality of the feature vector into a lower dimensional space by selecting the best subset features from the original feature set.

This paper is organized as follows. A related work is discussed in Section II. The main phases of the Ontology based dimensionality reduction for text documents classification model are introduced, proposing an efficient approach for dimensionality reduction of feature vector size based on the hierarchy of WordNet ontology, in Section III, Experimental results and performance evaluation are presented in Section IV. Finally, conclusions are given in Section V.

II. RELATED WORKS

In this section, we briefly review some background research including dimensionality reduction applied to document datasets, some previous attempts to apply semantic knowledge to enhance the classification accuracy.

In [7] [8] [9], A full Survey and a comparative study between Dimensionality Reduction Techniques for the Classification of text documents been introduced. It concentrates on the filter approach to achieve dimensionality

978-1-5090-5507-4/17/$31.00 ©2017 IEEE
ICIS 2017, May 24-26, 2017, Wuhan, China

reduction (DR), and proposed for DR technique to improved classification accuracy and a saving in the feature set size.

In [10], two feature selection techniques - Chi-Square and Information Gain Ratio and two feature extraction techniques – PCA and Latent Semantic Analysis are used for the analysis (LSA). It is found that feature extraction techniques offer better performance for the classification, give stable classification results with the different number of features chosen, and robustly keep the performance over time.

In [1] [2] [11] [12] [13] [14], a full review of the current trends for text documents classification is introduced. Also [15] shows that the Traditional text classification methods do not consider the semantic relationships among words so that cannot accurately represent the meaning of documents.

To overcome this problem, many previous works as in [3] [15] [16] [17] [18], were introducing the use of the semantic information from ontologies, such as WordNet ontology, to improve the accuracy of text mining tasks. WordNet is considered to be one of the largest and most widely used lexical databases of English [19]. Generally, it maps all the stemmed words from the text documents into their specific lexical categories in WordNet lexical database.

This paper proposes an approach to achieve classification of web text documents with semantic similarities in order to reduce the size of the used feature vector for the classification process without losing information in the text. The performance of the classification result has evaluated with the use of the F-Measure, Precision, recall, and the Classification accuracy.

III. ONTOLOGY BASED TEXT DOCUMENT CLSSIFICATION MODEL

The overall classification model passes through two stages; the Learning Stage, as explained in (III.A.), which consists of two main phases; feature extraction and the ontology based feature selection phases, and the Classification Stage, as explained in (III.B.), which consists of two main phases; weighting the feature vector and applying the classification algorithm to assign classes to test documents. The functional block diagram of the proposed Ontology based text documents classification model is depicted in Fig. 1.

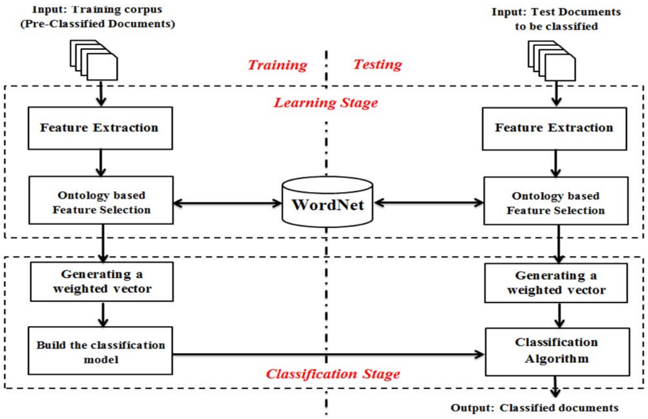

Fig. 1. Block diagram of the Ontology based text documents classification model.

A. *The learning stage*

The first issue that needs to be addressed in text classification is how to represent texts [20] [12]; not only to ease the manipulation of them by machines, saving the processing time and the used memory, but also to retain as much information as needed without any losses.

In order to represent the text documents; the Vector Space Model (VSM) is considered the most commonly used text representation technique, in which each document is represented as a vector ,Bag of Words (BoW), i.e., each document is represented by the set of words it contains and their frequency regardless their order [1] [16] [21] .

In our approach, we use a method in which we apply two main phases on both the training and the testing text documents to build the feature vector for mining tasks; the first is the Feature extraction phase, explained in (III.A.1.), and the second is the Ontology based Feature Selection Phase, explained in (III.A.2.).

1) Feature extraction phase : The feature extraction phase aims in preprocess the input documents; extracting the BoW that represents these documents. Algorithm 1, Provides a high-level pseudo-code description of used algorithm. TABLE I. contains the explanation of the basic elements of the algorithm.

ALGORITHM 1
FEATURE EXTRACTION METHOD

```
FeatureExtraction(TextDocument td,StoppingWords sw){
/* Document Parsing */
pd :=Parse (td)
/* BoW Extraction */
IBoW : = Extraxct_BoW(pd)
/* Reading words from the stopping list file */
SwL=Read_Words(sw)
/*stopping words removal*/
∀ $IBoW_i \in IBow${
        ∀ $SwL_j \in SwL${
                if Word_Compare($IBoW_i, SwL_j$) = False{
                        $MBoW := MBoW \cup IBoW_i$
                }
        }
}
/*Data Cleaning*/
RE="^[A-Za-z0-9 _]*[A-Za-z0-9][A-Za-z0-9 _]*$"
∀ $MBoW_i \in MBow${
        if PatternMatcher($MBoW_i$, RE)=True{
                $BoW := BoW \cup MBoW_i$
        }
}
/*Word Stemming*/
∀ $BoW_i \in Bow${
        $SBoW := SBoW \cup$ StemTerm($BoW_i$)
}
Return SBoW
}
```

TABLE I
BASIC ELEMENTS OF ALGORITHM 1

#	Element	Definition
1	td (TextDocument)	Text document which is in unstructured form.
2	sw (StoppingWords)	File which contains a list of stopping word, the auxiliary verbs, adverbs, etc.
3	pd (parsed document)	Text document which is parsed in a readable format suitable for the next steps.
4	SwL (Stopping Words List)	List of Stopping words.
5	IBoW (Initial BoW List)	List of BoW (Words extracted from the text that haven't been checked yet).
6	MBoW (Modified BoW List)	List of BoW (Words extracted from the text that after removing stopping words).
7	RE (Regular Expression)	The regular expression used to allow words which contain only English letters and numeric values without any symbols.
8	BoW (Cleaned BoW List)	List of BoW after removing non-English and symbolic words.

It consists of the following tasks:

a) Natural Language Processing (NLP) Parser: which is the responsible of processing text to detect sentences, tokens, by separating the words for the analysis. e.g., the word (can't) should be separated into two words (can) and (not) for a good further text analysis, and perform the Part-Of-Speech (PoS) tagging to mark up the words in a text as corresponding to a particular part of speech such as verbs, nouns, adjectives, etc. [22] .Fig. 2, provides a sample example of the parsing process, with the sentence "The Egyptian army is fighting terrorism." First, the words are marked as corresponding to a particular part of speech, by means of POS tagging, the tagged components are: (DT: Determiner, JJ: Adjective, NN: Common Noun, VBZ: Verb, 3rd person singular present, and VBG: Verb, gerund or present participle.),after applying the PoS tagging, Chunking is used to divide the text in syntactically correlated parts of words. In this example the result is only noun and verb phrases: (NP: Noun Phrase , VP: Verb Phrase) .

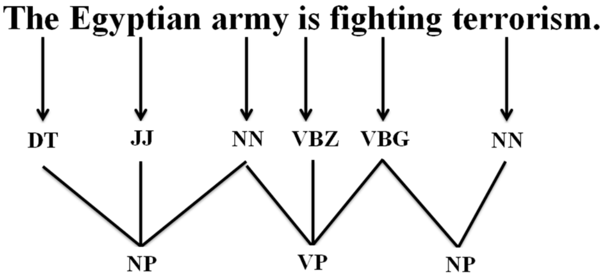

Fig. 2. Sample output of the NLP Parser

b) Stopping words removal: in which a removal of the stopping words, such as (a, an, in, at,…, etc.) ,also the auxiliary verbs, adverbs, etc., is done by searching for words in pre-existing Stopping words list [23].

c) Data Cleaning: in which regular expressions are used to Remove non English words and the words that contains symbolic characters; that accepts only the words that contain English characters and numbers and words with any combination of them, by using mechanism that detects takes only the words that matches the expression and discard those which don't match.

d) Stemming: it is a task in which each of the extracted words are replaced by its morphological root by applying any of the stemming algorithms. It is fundamental to avoid the redundancy of extracting the different equivalent morphological forms in which a word can be presented. TABLE II shows the results of stemming the set of words of the leftmost column.

TABLE II
SAMPLE RESULTS OF APPLYING PORTER STEMMING ALGORITHM

Words	Stemmed word
Fighting Fights	Fight
Concentrate Concentrates Concentrating	Concentr

*2) Ontology based Feature Selection Phase:*The ontology based feature Selection phase aims in reducing the dimensionality of the extracted BoW, based on the hierarchy of WordNet ontology to eliminate words that has no relation with any of WordNet lexical categories, without losing information on the text . ALGORITHM 2 provides a high-level pseudo-code description of the proposed algorithm. TABLE III contains the explanation of the basic elements of the algorithm.The main idea of Algorithm 2., is taking the benefit of the hierarchal structure of the WordNet ontology to remove any word in the stemmed BoW that has no path with any of the WordNet lexical categories by applying WuPalmer similarity measure as one of path length based measures of the semantic similarity measures [24] [25] is the function of path length and depth in path-based measures. Fig. 3, clarifies the idea behind the WuPalmer Similarity measure is defined as [26].

$$\mathrm{sim}_{wp}(c_1.c_2) = \frac{2 \times \mathrm{N}}{N_1 + N_2 + 2 \times \mathrm{N}} \qquad (1)$$

Where: simwp denotes the similarity value, C1 and C2, are two concepts. N, is the distance from the hierarchy root to the two concepts' common parent concept. N1and N2, are the positions of the concepts C1 and C2 in the taxonomy relatively to the position of the most specific common.

ALGORITHM 2
PROPOSED ONTOLOGY BASED FEATURE SELECTION METHOD

```
OntologBasedSelection(StringList BoW ,WordNetOntology wo){
/* load the WordNet LexicalDatabase */
db=Load_ Lexical_Database(wo)
LC=extract_categories(db)
/* selection process for dimensionality reduction */
∀ BoW_i ∈ Bow{
        Sim_Sum=0
        ∀ LC_j ∈ LC{
                Sim_Matrix[i][j]= sim(BoW_i, LC_j)
                Sim_Sum + =  Sim_Matrix[i][j]
        }
        if Sim_sum > 0{
                RBoW : = RBoW ∪ BoW_i
        }
}
Return RBoW
}
```

TABLE III
BASIC ELEMENTS OF ALGORITHM 2

#	Element	Definition
1	BoW (BoW List)	List of BoW out after applying Algorithm1
2	wo (WordNet Ontology)	The used WordNet 2.1 ontology
3	db (WordNet Lexical Database)	The WordNet Lexical Database.
4	*LC* (WordNet Lexical categories List)	List of the extracted WordNet lexical categories from the WordNet lexical database.
5	Sim_Sum (Similarity Sum)	A variable used to store the summation of the obtained similarity measure for each word in the BoW list, and been set to 0 in each iteration of new word.
6	Sim_Matrix (Similarity matrix)	Matrix contains the WuPalmer similarity measure between each word from the BoW list and each of the extracted WordNet Lexical categories in the LC list
7	RBoW (Reduced BoW List)	List of BoW after reduction.

B. The Classification stage

Mainly, the classification stage consists of two phases; the first is generating a weighted feature vector for each of the documents, either training or testing ones as explained in (III.B.1.), while the second concerns with the building a classification model and applying a classification algorithm, using some of the well-known classifiers included in WEKA datamining tool [27]such as Naive-Bayes, J48, JRip, and SVM as explained in (III.B.2.).

1) Generating a weighted feature vector: The task of generating a weighted feature vector, aims in assigning Weights to each word in the feature vector to give an indication of the importance of them.one of the most straight forward and useful techniques to weight words is the Term Frequency; the limitations of using this technique as it doesn't take the length of the documents into account [16]. However, [28] [29], introduces the use of the TFIDF as a weighting technique is a straight forward solution to this problem. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. It is composed by two terms; the first computes the normalized Term Frequency (TF), the second term is the Inverse Document Frequency (IDF).

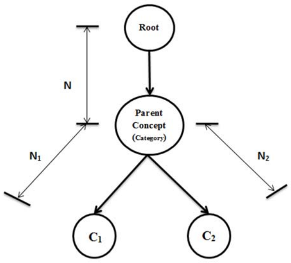

Fig. 3. The concept of the WuPalmer similarity measure

a) TF: This measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones, as follows [30]:

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}} \tag{2}$$

b) IDF: This measures how important a term is. While computing TF, all terms are considered equally important, as follows [30]:

$$IDF(t) = \frac{log(Total\ number\ of\ documents)}{Number\ of\ documents\ with\ term\ t\ in\ it} \tag{3}$$

c) Finally, the TFIDF measure: is calculated by the multiplication of TF and IDF, as follows [30]:

$$TFIDF(t) = TF(T) \times IDF(t) \tag{4}$$

This means that larger weights are assigned to terms that appear relatively rarely throughout the corpus, but very frequently in individual documents.

2) Building a classification model and applying the classification algorithm: Some of the well-known classifiers included in WEKA datamining tool such as Naive-Bayes, J48, JRip, and SVM have been used to perform testing of the proposed approach as explained in section IV.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we discuss our experimental setup and the results for evaluating the performance of our proposed approach. The experiments are divided into two parts. The first part works on reuters-21578 dataset [31] and evaluates the classification system when applying PCA as one of the traditional feature reduction techniques. The second part works also on reuters-21578 dataset but using the help of WordNet ontology for the dimensionality reduction process. Then we compare our proposed ontology based approach against the PCA ones applying different classifiers on WEKA.

A. Dataset Used

The Reuters-21578 dataset has been used in many text categorization experiments; the data was collected by the Carnegie group from the Reuters newswires in 1987. It consists of 21578 collections of new stories classified into topics. However, not all documents have a topic, there exist some of the documents that have more than one topic, and not all documents have a text in the news body.

So, we used only documents that have only one topic and have text in its body, ignoring those which have no text in the body and associated with more than one topic. The dataset has classes of different sizes; some classes have large size such as earn class that has 3734 documents belonging to it, while other classes have size less than 5 documents such as rice. So, in our experiments, we used a reduced, unbiased subset from this corpus as our dataset for training and testing each group contains between 30 and 100 documents as indicated in Table IV.

TABLE IV

NUMBER FOR TRAINING AND TESTING DOCUMENTS USED

Category	#Training	#Test	Total
gold	70	20	90
money-supply	70	17	87
gnp	49	14	63
cpi	45	15	60
cocoa	41	12	53
alum	29	16	45
grain	38	7	45
copper	31	13	44
jobs	32	10	42
reserves	30	8	38
rubber	29	9	38
iron-steel	26	11	37
ipi	27	9	36
nat-gas	22	11	33
veg-oil	19	11	30
TOTAL	558	183	741

B. Evaluation criteria

All documents for training and testing Passes through the stages in section III, Experimental results reported in this section are based on precision, recall and F1 measures.

The F1 measure is the harmonic mean of precision and recall as follows [32]:

$$\mathrm{F}_1(\mathrm{recall.\,precision}) = \frac{2\times\mathrm{recall}\times\mathrm{precision}}{\mathrm{recall}+\mathrm{precision}} \tag{5}$$

In the above formula, precision and recall [32] are two standard measures widely used in text categorization literature to evaluate the algorithm's effectiveness on a given category where:

$$\mathrm{precision} = \frac{\text{true positive}}{\text{true positive+false positive}} \tag{6}$$

$$\mathrm{recall} = \frac{\text{true positive}}{\text{true positive+false negative}} \tag{7}$$

$$\mathrm{Accuracy} = \frac{\text{\# of Correctly classifieddocuments}}{\text{total number of documents}} \tag{8}$$

C. Experimental results

To test the ontology-based feature reduction approach and compare it with the traditional approach (PCA), an experiment was implemented on the reuters-21578 dataset, as in (IV.A.), using both approaches. Table V, shows a comparison between the ontology-based and PCA-based feature reduction approaches. It could be noticed that the PCA-based feature reduction approach is superior to the ontology-based feature reduction approach.

TABLE V

COMPARISON BETWEEN THE ONTOLOGY-BASED AND PCA-BASED APPROACHES

Dimensionality reduction Technique	Feature vector size	Reduced vector size	Reduction percentage
PCA	4006	512	87%
Ontology based		1804	55%

But Table VI, shows the resultant performance (accuracy, F-measure, precision, and recall) of the PCA-based Web text document classification and ontology-based Web text document classification using four classifiers (Naive-Bayes, JRip, J48, and SVM). It could be noticed that, the accuracy of J48 classifier after applying the PCA for feature reduction (49.54%) is much better than the other classifiers. Also, It could be noticed that, the accuracy of SVM classifier after applying the ontology for feature reduction (85.13%) is much better than the other classifiers.

In conclusion, the ontology-based Web text document classification is superior to the PCA-based Web text document classification using four classifiers as the reduction process using PCA depends on statistical methods used to reduce the number of features in the feature vector by lumping highly correlated features together [6] [33] regardless the importance of the meaning of these features , which may results in the existence of one or more features that are important for describing the meaning of a particular document would be eliminated when applying the PCA reduction technique, and making the reduced vector loses some important information.

TABLE VI
EVALUATION MEASURES USING FOUR DIFFERENT CLASSIFIERS (NAIVE-BAYES, JRIP, J48, AND SVM)

Category	Performance measure	PCA-based feature reduction	Ontology-based feature reduction
Naive-Bayes	Accuracy (%)	**27.9279**	**75.2252**
	F-Measure	**0.284**	**0.759**
	Recall	**0.279**	**0.752**
	Precision	**0.406**	**0.798**
JRip	Accuracy (%)	46.3964	75.6767
	F-Measure	**0.458**	**0.762**
	Recall	**0.464**	**0.757**
	Precision	**0.529**	**0.795**
J48	Accuracy (%)	49.5495	81.0811
	F-Measure	**0.498**	**0.814**
	Recall	**0.495**	**0.811**
	Precision	**0.539**	**0.843**
SVM	Accuracy (%)	**32.4324**	85.1351
	F-Measure	**0.323**	**0.85**
	Recall	**0.324**	**0.851**
	Precision	**0.607**	**0.876**

V. CONCLUSION

In this paper, we proposed an efficient technique for reducing the dimensionality of the used feature vector for web text documents classification. We performed an experimental evaluation on Reuters-21578 dataset and compared the proposed ontology based dimensionality reduction technique against the PCA as one of the corresponding classical dimensionality reduction methods.

However, there are still some limitations in our approach as some important words which are not included in WordNet lexicon will not be considered and will be discarded. In addition, the proposed method concerns only with the dimensionality reduction process.

In future work, we recommend extending this work by utilizing the WordNet ontology to measure the semantic similarity between documents instead of using the traditional TFIDF technique to construct the weighting building the feature to perform ontology based text document classification

REFERENCES

[1] R. M. A. J. Rajni Jindal, "Techniques for text classification: Literature review and current trends," *Webology,* vol. 12, no. 2, p. Article 139., 2015.

[2] B. B. B. K. K. Aurangzeb Khan, "An Overview of E-Documents Classification," in *International Conference on Machine Learning and Computing*, Singapore, 2011.

[3] a. T. G. Kerem Çelik, "A Comprehensive Analysis of using Semantic Information in Text Categorization," in *The IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA 2013)*, Albena, 2013.

[4] a. M. M. Pradnya Kumbhar, "A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification," *International Journal of Science and Research (IJSR) ,* vol. 5, no. 5, pp. 1267-1275, 2016.

[5] M. W. Mwadulo, "A Review on Feature Selection Methods For Classification Tasks," *International Journal of Computer Applications Technology and Research,* vol. 5, no. 6, pp. 395-402, 2016.

[6] a. B. Y. Tonglin Zhang, "Big Data Dimension Reduction using PCA," in *IEEE International Conference on Smart Cloud*, New York, 2016.

[7] a. S. M. G. Swati Kaur, "A Survey on Dimension Reduction Techniques for Classification of Multidimensional Data," *International Journal of Science Technology & Engineering (IJSTE),* vol. 2, no. 12, pp. 31-37, 2016.

[8] J. L. ,. Q. Z. a. Y. W. Haozhe Xie, "Comparison among dimensionality reduction techniques based on Random Projection for cancer classification," *Computational Biology and Chemistr,* vol. 65, p. 65–172, 2016.

[9] D. A. Said, "Dimensionality Reduction Techniques For Enhancing Automatic Text Categorization," *Master Thesis,Cairo University,* 2007.

[10] a. S. K. R. Masoumeh Zareapoor, "Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection," *international journal Information Engineering and Electronic Business,* vol. 2, pp. 60-65, 2015.

[11] a. D. V. P. Sayali Rasane, "Handling Various Issues In Text Classification : A Review," *International Journal on EmergingTrends in Technology (IJETT),* vol. 3, no. 1, pp. 4076-4082, 2016.

[12] S. a. C. C.Uma, "A Survey Paper on Text Mining Techniques," *International Journal of Engineering Trends and Technology (IJETT),* vol. 40, no. 4, pp. 225-229, 2016.

[13] L. P. a. N. M. N. Venkata Sailaja, "Survey of Text Mining Techniques, Challenges and their Applications (IJCA)," *International Journal of Computer Applications,* vol. 146, no. 11, pp. 30-35, 2016.

[14] G. P. D. U. K. J. Omkar Ardhapure, "Comparative Study Of Classification Algorithm For Text Based Categorization," *International Journal of Research in Engineering and Technology (IJRET),* vol. 5, no. 2, pp. 217-220, 2016.

[15] A. R. a. M. A. B. Zakaria Elberrich, "Using WordNet for Text Categorization," *The International Arab Journal of Information Technology,* vol. 5, no. 1, pp. 16-24, 2008.

[16] a. D. K. Julian Sedding, "WordNet-based Text Document Clustering," in *ROMAND '04 Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data*, Geneva, 2004.

[17] a. H. C. T. Wei, "Measuring Word Semantic Relatedness Using WordNet-Based Approach," *Journal of Computers,* vol. 10, no. 4, pp. 252-259, 2015.

[18] Y. L. H. C. Q. Z. a. X. B. T. Wei, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Systems with Applications,* vol. 42, no. 4, p. 2264–2275, 2015.

[19] C. C. S. M. ,. P.-J. C. Samia Iltache, "Using Domain Ontologies for Classification and Semantic Interpretation of Documents," in *The Second International Conference on Big Data, Small Data, Linked Data and Open Data*, Portugal, 2016.

[20] D. S. G. S. M. B S Harish, "Representation and Classification of Text Documents: A Brief Review," *IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition",* 2010.

[21] P. X. a. C. B., "Document Classifications Based on Word Semantic Hierarchies (IASTED)," *the International Conference on Artificial Intelligence and Applications,* vol. 5, pp. 362-367, 2005.

[22] "Penn Part of Speech Tags," Computer Science Department at New York University, [Online]. Available: https://cs.nyu.edu/grishman/jet/guide/PennPOS.html. [Accessed 10 1 2017].

[23] "Onix Text Retrieval Toolkit," Lextek International, [Online]. Available: http://www.lextek.com/manuals/onix/stopwords1.html. [Accessed 7 january 2017].

[24] R. H. J. G. Lingling Meng, "A Review of Semantic Similarity Measures in WordNet," *International Journal of Hybrid Information Technology,* vol. 6, no. 1, pp. 1-12, 2013.

[25] D. J. A. K. M. S. Anitha Elavarasi, "A Survey on Semantic Similarity Measure," *International Journal of Research in Advent Technology,* vol. 2, no. 3, pp. 389-398, 2014.

[26] Z. W. a. M. Palmer., "Verb semantics and lexical selection," in *In Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics*, New Mexico, 1994.

[27] M. L. Group, "Downloading and installing Weka," The University of Waikato, [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/downloading.html. [Accessed 10 January 2017].

[28] a. J. W. C. Yang, "Text Categorization Based on a Similarity Approach," in *In Proceedings of International Conference on Intelligent System and Knowledge Engineering*, China, 2007.

[29] H. C. L. R. W. P. W. K. F. &. K. K. L. Wu, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS),* vol. 26, no. 3, p. 13, 2008.

[30] P. R. ,. a. H. S. Christopher D. Manning, Introduction to Information Retrieval, Cambridge : Cambridge University Press, 2008.

[31] "The Reuters dataset is available to be downloaded in sgml format from," [Online]. Available: http://www.daviddlewis.com/ressources/testcollections/reuters21578/. [Accessed 12 January 2017].

[32] a. G. L. Marina Sokolova, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management,* vol. 45, no. 4, p. 427–437, 2009.

[33] L. I. Smith, "A tutorial on Principal Components Analysis," International Institute of Information Technology, India, 2002.